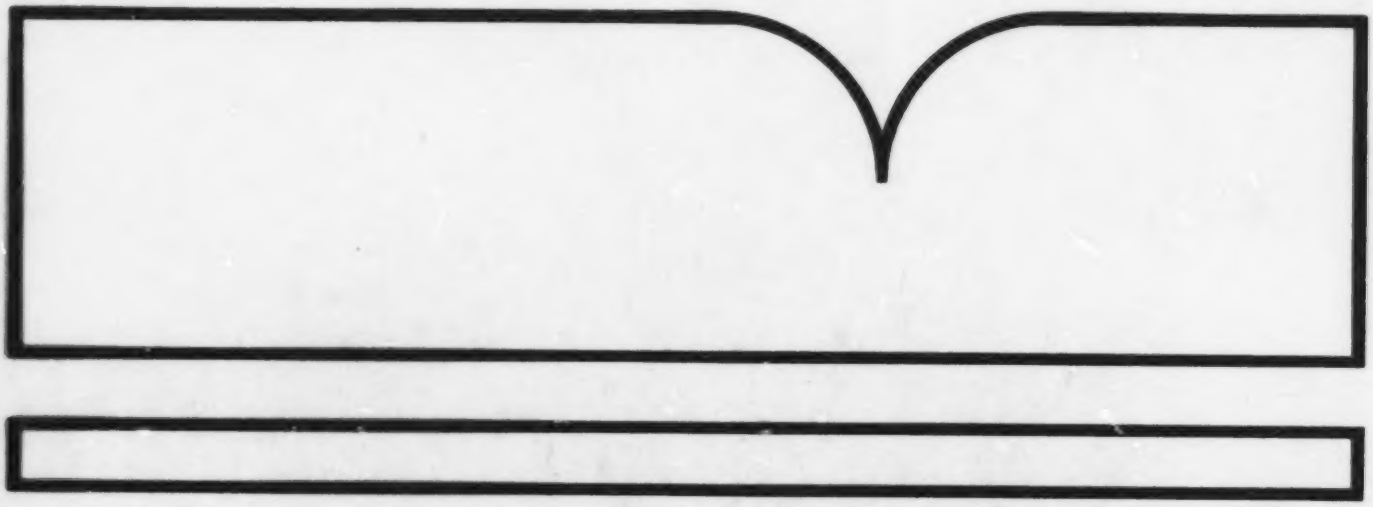
PB81-231516

Analysis of Coarsely Grouped Data from the Lognormal Distribution

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

Jun 80



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA (Please read Instructions on the reverse before completing)		MS-78-099
1. REPORT NO. EPA-600/J-80-299	2. JOURNAL ARTICLE	3. RE PB81-231516
4. TITLE AND SUBTITLE Analysis of Coarsely Grouped Data From the Lognormal Distribution		5. REPORT DATE July 1981
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Victor Hasselblad, Andrew G. Stead, Warren Galke		8. PERFORMING ORGANIZATION REPORT
9. PERFORMING ORGANIZATION NAME AND ADDRESS Biometry Division Research Triangle Park, N.C, 27711		10. PROGRAM ELEMENT NO. A9RA1B
		11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Health Effects Research Laboratory U.S. Environmental Protection Agency Research Triangle Park, N.C. 27711		13. TYPE OF REPORT AND PERIOD COVERE
		14. SPONSORING AGENCY CODE EPA-600/11

15. SUPPLEMENTARY NOTES

Published In: Journal of the American Statistical Assn. 75(372): 771-778, 1980

16. ABSTRACT

A missing information technique is applied to blood lead data that is both grouped and assumed to be lognormally distributed. These maximum likelihood techniques are extened from the simple lognormal case to obtain solutions for a general linear model case. Various models are fitted to the data, and likelihood ratio statistics are computed to test for significance of various parameterizations. The techniques are applied to a data set of over 130,000 blood lead values collected by the city of New York. The data collected from 1970 to 1976, are part of a large-scale screening program and have implications for the current ambient air lead standard.

17. KEY WORDS AND DOCUMENT ANALYSIS		
a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

PRODUCT OF:
NATIONAL TECHNICAL INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
RELEASE TO PUBLIC	UNCLASSIFIED	
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA-600/J-80-299
Journal Article

Analysis of Coarsely Grouped Data From the Lognormal Distribution

VICTOR HASSELBLAD, ANDREW G. STEAD, and WARREN GALKE*

A missing information technique is applied to blood lead data that is both grouped and assumed to be lognormally distributed. These maximum likelihood techniques are extended from the simple lognormal case to obtain solutions for a general linear model case. Various models are fitted to the data, and likelihood ratio statistics are computed to test for significance of various parameterizations. The techniques are applied to a data set of over 130,000 blood lead values collected by the city of New York. The data, collected from 1970 to 1976, are part of a large-scale screening program and have implications for the current ambient air lead standard.

KEY WORDS: Missing information; Grouped lognormal data; Blood lead; Maximum likelihood; EM algorithm.

1. INTRODUCTION

Although there were early reports of lead poisoning in Baltimore during the 1930's, it was not recognized as a major public health problem in the United States until the 1960's. An alarming number of people, particularly young children, were found to have dangerously high blood lead burdens. Affected children lived primarily in core urban areas and in industrial areas close to lead smelters. To address the problem, several metropolitan areas initiated public health surveillance programs.

New York City began large-scale screening of children for elevated blood lead levels in March 1970 (Guinee 1971). By December 1976 over 340,000 individual blood lead determinations had been obtained through the program. At this time, the U.S. Environmental Protection Agency (EPA) was in the process of setting an ambient air quality standard for lead. The New York City data set was considered valuable for the following reasons: (a) It contained more than 100 times the number of blood lead values available from any other data set, (b) it contained data on young children in the core urban area, who were considered to be a high risk group, and (c) it contained longitudinal data on a well-defined population during a period in which the lead in gasoline was being decreased. Unfortunately, data on the actual amounts of lead in New York City gasoline were not available.

Analysis of the New York City data set presented an interesting problem. Since the data set was collected for screening purposes, the values were recorded in broad intervals. Before January 1, 1975, the intervals (in μg/100 ml) were 1–14, 15–24, 25–34, and so on. Subsequently, the intervals were changed to 1–19, 20–29, 30–39, and so on. In a few cases, subgroups contained over 50 percent of their values in the initial interval.

Several authors, including Yankel, von Linern, and Walter (1977), Tepper and Levin (1975), Azar, Snee, and Habibi (1975), Hasselblad and Nelson (1975), and Billick, Curran, and Shier (1979) have found that the lognormal distribution provides a good fit to blood lead data from relatively homogenous groups. Billick, Curran, and Shier (1979) previously analyzed the New York City data by representing each child's blood lead level reading by the midpoint of the interval in which it was coded. They assumed a lognormal distribution and, after taking logarithms of the midpoints, used standard regression analyses.

The use of midpoints created two potential problems. First, the change of the interval definition in 1975 could create an inconsistent bias in the estimate of the annual trends. Second, the coarseness of the grouping could bias the estimate of the geometric standard deviation.

The estimate of the geometric standard deviation was particularly important. In setting an ambient air lead standard, EPA (1978) argued that a population blood lead mean of 15 μg/100 ml would leave only a small fraction of the population above a level of 30 μg/100 ml (a blood lead level over 30 μg/100 ml is generally considered to be an unacceptable medical risk). If the geometric standard deviation is 1.4, the fraction above 30 μg/100 ml would be approximately 1.97 percent. If the geometric standard deviation is 1.5, the fraction above 30 μg/100 ml would be 4.37 percent. The difference in these two fractions is especially large when they are multiplied by the hundreds of thousands of urban children who are potentially at risk.

Our assumption in the analysis was that the data originated from a lognormal distribution where the

* Victor Hasselblad and Andrew G. Stead are Statisticians with the Health Effects Research Laboratory, and Warren Galke is an Epidemiologist with the Environmental Criteria Assessment Office, U.S. Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC 27711. The authors would especially like to thank Irwin H. Billick, Anita Curran, and Douglas Shier for their cooperation in providing the data base and for their helpful comments as reviewers; William Nelson and John Creason for their poignant comments and suggestions as reviewers; and Barbara Crabtree for her time and patience in typing this article. The editors and referees made many excellent suggestions for which we are also grateful.

JUN 16 1981

1. Number of Usable Analyzed Blood Samples by Ethnic Group, Age, and Year of Collection

Ethnic Group	Age in Years	Year of Sample Collection							Totals
		1970	1971	1972	1973	1974	1975	1976	
Blacks	≤1	302	1,137	931	540	252	376	342	3,880
	1–2	1,379	4,077	3,295	1,741	829	1,087	939	13,347
	2–3	1,290	3,090	2,610	1,496	694	853	678	10,711
	3–4	1,338	3,272	2,489	1,624	713	655	512	10,603
	4–5	1,688	3,669	2,442	1,462	610	508	448	10,827
	5–6	1,444	3,526	2,044	1,218	492	344	303	9,371
	≥6	1,411	4,498	2,965	1,676	590	797	651	12,588
Whites	≤1	56	136	89	43	34	45	35	438
	1–2	236	547	367	184	152	154	123	1,763
	2–3	240	387	250	135	106	110	87	1,315
	3–4	194	364	211	120	76	90	58	1,113
	4–5	220	414	170	110	83	66	81	1,144
	5–6	193	456	122	113	62	62	40	1,048
	≥6	146	497	146	121	91	136	90	1,227
Puerto Rican	≤1	336	871	607	317	139	183	169	2,622
	1–2	1,486	3,685	2,333	1,318	498	630	555	10,505
	2–3	1,388	2,861	1,707	1,106	398	416	371	8,247
	3–4	1,366	2,625	1,476	1,067	433	368	303	7,638
	4–5	1,344	2,651	1,507	1,183	366	314	300	7,665
	5–6	1,162	2,573	1,185	913	289	223	187	6,532
	≥6	1,177	3,499	1,364	996	396	506	343	8,281
Other	≤1	83	51	28	17	11	17	38	245
	1–2	282	231	77	105	51	66	70	882
	2–3	279	159	52	103	28	42	42	705
	3–4	250	138	43	59	33	38	48	609
	4–5	285	144	29	73	30	29	30	602
	5–6	254	165	32	55	37	32	25	600
	≥6	250	138	32	32	33	52	48	585
Totals		20,079	45,861	28,603	17,927	7,526	8,199	6,916	135,111

geometric mean was a function of factors such as age, race, and year of collection. To handle the parameter estimation problem, we turned to a technique developed by Orchard and Woodbury (1972) and Dempster, Laird, and Rubin (1977). Known as the missing information principle (MIP) or the EM Algorithm, this procedure almost always converges, and on a large scale computer, such as the Univac 1110, the computation time is relatively short.

2. DESCRIPTION OF DATA SET

Since January 1970 the New York City Bureau of Lead Poisoning has maintained computerized records of all children screened by venipuncture for elevated blood lead levels. Each record contains the child's birthdate, ethnic group, screening date, a geographic location indicator, and the sample blood lead measurement. Unfortunately, duplicate measures of each sample were not available to investigate the variability of the analytic method as Lucas (1978) has done for other blood lead data sets.

The data base for children screened from March 1970 to December 1976 consisted of 344,512 records. Duplicating the editing procedures outlined by Billick, Curran, and Shier (1979), we restricted our analyses to the 178,506 records classified as initial screens. (For a more detailed description of the data, the reader is referred to the paper by these authors.) Table 1 presents the number of initial screening tests by year, age, and ethnic group.

3. STATISTICAL CONSIDERATIONS

The basic assumption for the New York City blood lead data set is that the original values come from a lognormal distribution. The mean of this distribution depends on several independent variables such as age, race, and year of collection. Using dummy variables to indicate the levels of these independent variables, the logarithms of the blood lead values are normally distributed with mean μ_i and common variance σ^2, where

$$\mu_i = \sum_{k=1}^{K} \beta_k X_{ik}\,, \quad i = 1, 2, \ldots, n \tag{3.1}$$

and i denotes the various combinations of the independent variables. The β_k's are unknown parameters to be estimated. The observations at a particular level, i, have been grouped into j intervals (c_j, c_{j+1}), $j = 1, 2, \ldots, J$, so that only the frequencies, n_{ij}, of observations occurring in these intervals are available. The total number of observations at a particular level i is $n_{i.} = \sum_{j=1}^{J} n_{ij}$. By replacing each c_j with its logarithm, the distribution can be transformed to a normal one. The problem is to estimate $\beta_1, \beta_2, \ldots, \beta_k$, and σ^2, and

2. Counts of Blood Lead Levels by Year, Race, and Age

Year	Race	Age	(a) Blood Lead (micrograms/100 ml)						
			0–14	15–24	25–34	35–44	45–54	55–64	65+
1970	W	1–12	7	29	14	5	1	0	0
1970	W	13–24	24	85	85	31	6	4	1
1970	W	25–36	16	91	80	37	6	4	6
1970	W	37–48	13	83	56	26	11	5	0
1970	W	49–60	20	88	83	20	5	2	2
1970	W	61–72	14	76	71	24	5	1	2
1970	W	73+	17	70	35	14	7	0	3
1970	B	1–12	19	91	103	66	15	3	5
1970	B	13–24	53	313	487	321	126	46	33
1970	B	25–36	40	283	434	327	125	46	35
1970	B	37–48	49	313	491	312	114	37	22
1970	B	49–60	57	424	699	343	102	46	17
1970	B	61–72	58	410	585	274	84	27	6
1970	B	73+	41	411	625	227	85	15	7
1970	P	1–12	49	131	110	26	16	3	1
1970	P	13–24	117	543	550	169	64	21	22
1970	P	25–36	78	523	503	181	63	22	18
1970	P	37–48	87	491	492	215	48	21	12
1970	P	49–60	85	511	497	178	42	24	7
1970	P	61–72	77	440	437	154	42	8	4
1970	P	73+	66	509	407	151	31	10	3
1971	W	1–12	22	85	22	4	0	3	0
1971	W	13–24	63	271	145	44	18	3	3
1971	W	25–36	41	191	111	29	5	6	4
1971	W	37–48	42	172	108	31	9	2	0
1971	W	49–60	28	190	154	34	8	0	0
1971	W	61–72	50	250	126	23	6	0	1
1971	W	73+	61	260	143	27	5	0	1
1971	B	1–12	106	445	372	162	37	9	6
1971	B	13–24	120	1,061	1,561	819	307	113	96
1971	B	25–36	55	764	1,177	706	263	80	45
1971	B	37–48	65	790	1,338	760	236	61	22
1971	B	49–60	94	1,064	1,512	697	220	55	27
1971	B	61–72	112	1,135	1,427	640	162	35	15
1971	B	73+	137	1,571	1,836	738	174	31	11
1971	P	1–12	174	448	188	48	8	3	2
1971	P	13–24	407	1,774	1,065	299	101	25	14
1971	P	25–36	180	1,260	936	353	86	19	27
1971	P	37–48	165	1,150	917	300	64	15	14
1971	P	49–60	180	1,219	903	251	77	12	9
1971	P	61–72	187	1,278	819	208	56	16	9
1971	P	73+	261	1,726	1,114	280	96	10	12
1972	W	1–12	24	49	14	2	0	0	0
1972	W	13–24	70	196	78	14	6	2	1
1972	W	25–36	33	129	70	14	3	0	1
1972	W	37–48	31	110	58	9	3	0	0
1972	W	49–60	20	98	38	12	2	0	0
1972	W	61–72	11	77	24	10	0	0	0
1972	W	73+	25	79	30	12	0	0	0
1972	B	1–12	114	439	272	77	23	3	3
1972	B	13–24	150	1,306	1,140	455	159	44	41
1972	B	25–36	108	971	973	375	113	49	21
1972	B	37–48	112	1,002	929	339	78	18	11
1972	B	49–60	119	1,052	881	308	70	6	6
1972	B	61–72	136	903	725	224	45	10	1
1972	B	73+	203	1,427	1,004	262	57	8	4
1972	P	1–12	148	322	104	26	4	1	2
1972	P	13–24	392	1,234	543	113	28	14	9
1972	P	25–36	194	884	464	119	30	8	8
1972	P	37–48	133	804	383	111	34	6	5
1972	P	49–60	151	839	379	105	22	4	7
1972	P	61–72	136	622	318	83	18	4	4
1972	P	73+	186	730	360	70	14	2	2
1973	W	1–12	9	23	9	2	0	0	0
1973	W	13–24	31	91	45	10	4	1	2
1973	W	25–36	17	82	18	14	2	1	1
1973	W	37–48	10	73	25	9	2	0	1
1973	W	49–60	14	67	24	4	1	0	0
1973	W	61–72	19	61	27	4	2	0	0
1973	W	73+	13	76	27	5	0	0	0
1973	B	1–12	65	252	145	57	14	4	3

2. (Continued)

Year	Race	Age	(a) Blood Lead (micrograms/100 ml) 0–14	15–24	25–34	35–44	45–54	55–64	65+
1973	B	13–24	97	644	607	252	83	30	28
1973	B	25–36	90	561	518	214	76	22	15
1973	B	37–48	92	651	580	203	71	19	8
1973	B	49–60	92	662	465	184	47	7	5
1973	B	61–72	72	576	390	126	36	9	9
1973	B	73+	144	796	522	162	41	10	1
1973	P	1–12	70	157	61	22	6	0	1
1973	P	13–24	213	665	303	84	32	13	8
1973	P	25–36	132	542	308	87	14	17	6
1973	P	37–48	103	525	304	102	20	6	7
1973	P	49–60	145	598	317	92	26	2	3
1973	P	61–72	118	493	223	58	19	1	1
1973	P	73+	133	529	234	76	19	4	1
1974	W	1–12	8	21	3	2	0	0	0
1974	W	13–24	29	71	38	11	2	0	1
1974	W	25–36	15	49	30	11	1	0	0
1974	W	37–48	10	41	19	6	0	0	0
1974	W	49–60	10	50	17	2	3	1	0
1974	W	61–72	8	32	21	1	0	0	0
1974	W	73+	23	43	17	8	0	0	0
1974	B	1–12	27	132	65	23	5	0	0
1974	B	13–24	59	336	259	113	38	11	13
1974	B	25–36	57	257	230	99	38	10	3
1974	B	37–48	70	301	218	88	20	11	5
1974	B	49–60	45	302	187	51	20	3	2
1974	B	61–72	54	266	133	30	8	1	0
1974	B	73+	65	316	157	39	11	1	1
1974	P	1–12	27	71	32	6	3	0	0
1974	P	13–24	86	222	130	39	17	4	0
1974	P	25–36	41	190	117	34	10	4	2
1974	P	37–48	49	204	140	29	9	2	0
1974	P	49–60	42	190	100	27	6	1	0
1974	P	61–72	46	162	61	12	6	1	1
1974	P	73+	64	220	92	14	4	2	0

Year	Race	Age	(b) Blood Lead (micrograms/100 ml) 0–19	20–29	30–39	40–49	50–59	60–69	70+
1975	W	1–12	27	15	3	0	0	0	0
1975	W	13–24	79	54	19	2	0	0	0
1975	W	25–36	63	36	11	0	0	0	0
1975	W	37–48	58	26	5	1	0	0	0
1975	W	49–60	35	26	5	0	0	0	0
1975	W	61–72	42	17	3	0	0	0	0
1975	W	73–84	102	30	3	1	0	0	0
1975	B	1–12	186	135	48	5	1	1	0
1975	B	13–24	375	484	174	37	15	0	2
1975	B	25–36	292	387	130	34	6	3	1
1975	B	37–48	237	293	101	21	3	0	0
1975	B	49–60	200	229	71	6	2	0	0
1975	B	61–72	142	139	55	8	0	0	0
1975	B	73–84	442	273	71	9	2	0	0
1975	P	1–12	128	44	11	0	0	0	0
1975	P	13–24	355	206	53	10	3	1	2
1975	P	25–36	194	166	45	8	2	1	0
1975	P	37–48	173	143	44	7	1	0	0
1975	P	49–60	144	134	33	3	0	0	0
1975	P	61–72	116	83	19	4	1	0	0
1975	P	73–84	330	139	30	6	1	0	0
1976	W	1–12	25	9	0	1	0	0	0
1976	W	13–24	94	25	2	2	0	0	0
1976	W	25–36	51	28	5	2	0	0	1
1976	W	37–48	34	18	5	1	0	0	0
1976	W	49–60	56	20	4	1	0	0	0
1976	W	61–72	27	11	2	0	0	0	0
1976	W	73–84	64	23	3	0	0	0	0
1976	B	1–12	227	97	14	3	0	0	1
1976	B	13–24	486	305	99	38	8	2	1
1976	B	25–36	294	290	76	17	0	0	1
1976	B	37–48	240	197	64	8	1	1	1
1976	B	49–60	221	177	43	7	0	0	0

2. (Continued)

			(b) Blood Lead (micrograms/100 ml)						
Year	Race	Age	0–19	20–29	30–39	40–49	50–59	60–69	70+
1976	B	61–72	182	91	25	5	0	0	0
1976	B	73–84	388	191	63	8	1	0	0
1976	P	1–12	127	30	9	1	2	0	0
1976	P	13–24	358	149	39	4	4	0	1
1976	P	25–36	221	111	33	3	3	0	0
1976	P	37–48	173	105	20	5	0	0	0
1976	P	49–60	177	101	18	2	1	1	0
1976	P	61–72	129	39	17	2	0	0	0
1976	P	73–84	221	99	20	3	0	0	0

to make tests of hypotheses about the β's, given the design variables X_{ij}.

In the simple case in which $\mu_i \equiv \mu$, equations have been given by several authors. Perhaps the first of these was Gjeddebaek (1949), who also gave the estimates of the asymptotic variance-covariance matrix of the two parameters. Dempster, Laird, and Rubin (1977) were the first to point out that the EM algorithm gives a relatively simple method of solving the problem. They gave general equations for the problem, which included grouping, censoring, and truncation as special cases.

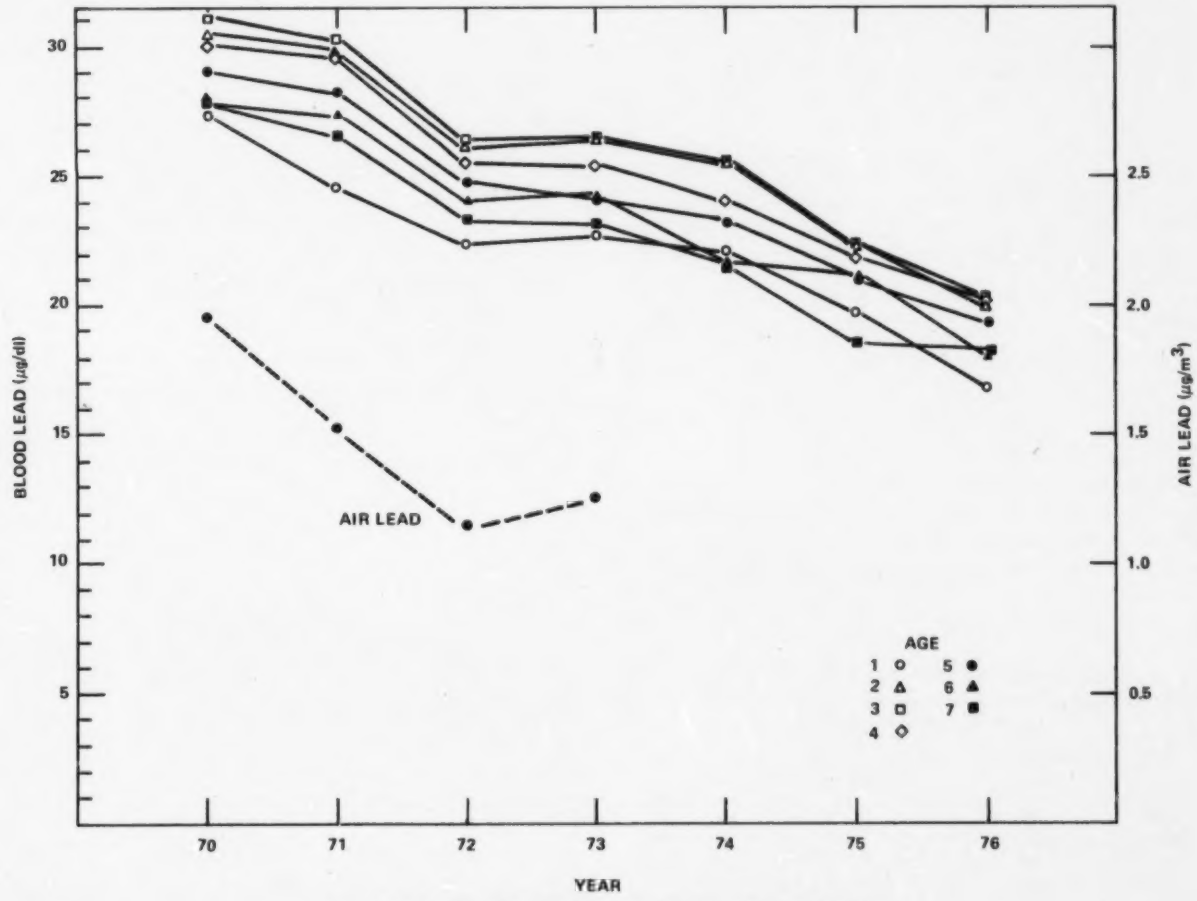
The extension to the general linear model problem is relatively straightforward. If we let y be the unobserved normal random variable such that

$$y_i = \mu_i + \epsilon_i \,, \tag{3.2}$$

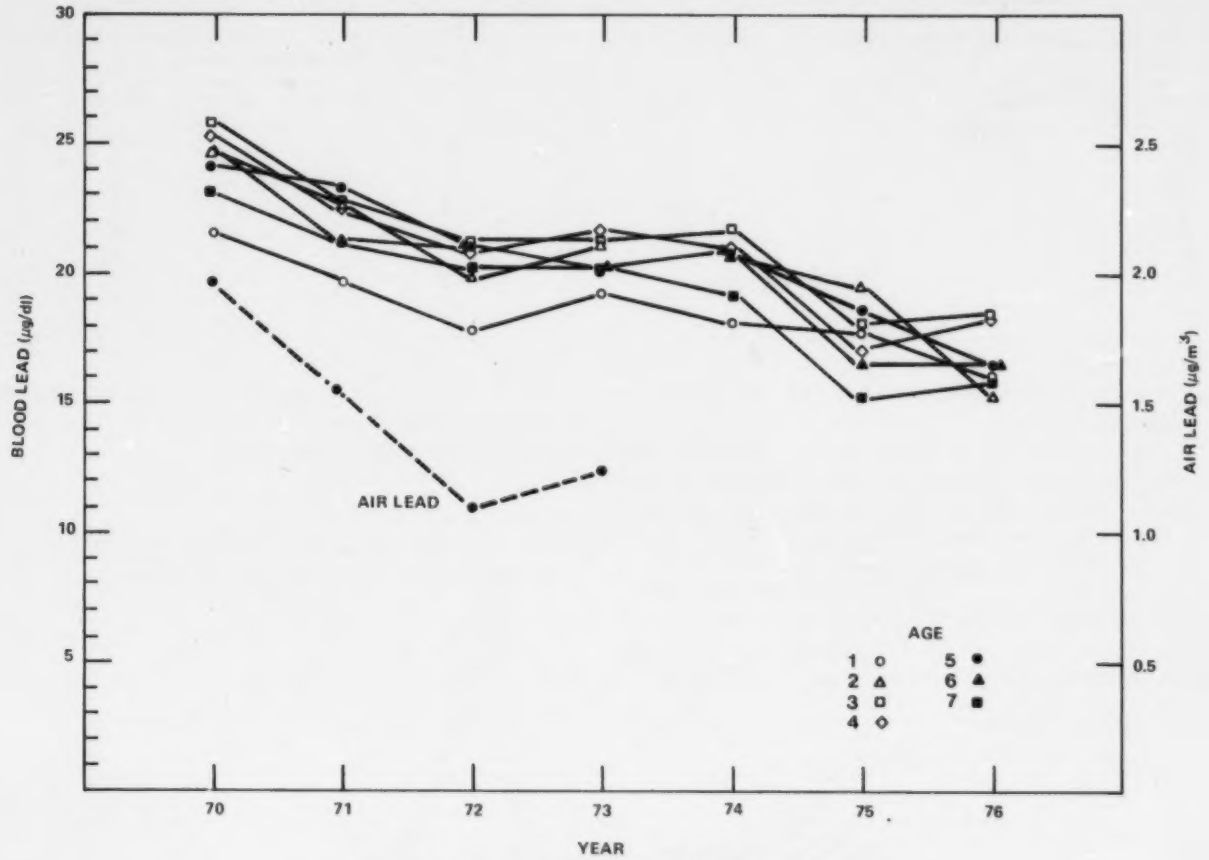
where ϵ_i is $N(0, \sigma^2)$, then the missing information (i.e., unknown sufficient statistics) can be estimated

$$\hat{E}(y_i | c_j < y_i < c_{j+1}) = \mu_i - z_{1i}\sigma^2 \tag{3.3}$$

$$\hat{E}(y_i^2 | c_j < y_i < c_{j+1}) = \sigma^2(1 + z_{2i}) + \mu^2 - \mu z_{1i}\sigma^2 \tag{3.4}$$



A. New York City Mean Blood Levels for Blacks



B. New York City Mean Blood Levels for Whites

where

$$z_{1i} = [f(c_{i+1}) - f(c_i)]/F_i \tag{3.5}$$

$$z_{2i} = -[c_{i+1}f(c_{i+1}) - c_iF(c_i)]/F_i \tag{3.6}$$

and

$$f(x) = (2\pi)^{-1} \exp[-(x-\mu)^2/2\sigma^2] \tag{3.7}$$

$$F_i = \int_{c_i}^{c_{i+1}} f(x)dx \ . \tag{3.8}$$

By using previous estimates of the parameters β_1, β_2, ..., β_k, and σ^2, along with (3.1), (3.5), (3.6), (3.7), and (3.8), the left-hand sides of (3.3) and (3.4) can be estimated for each interval (c_j, c_{j+1}) for each vector of observations i, $i = 1, 2, \ldots, N$. This defines the "E" step of the EM algorithm.

3. Air Lead Values From New York City[a]

	Manhattan (five-site average)			
Year	Winter	Spring	Summer	Fall
1969	—	—	1.43	1.64
1970	1.70	1.79	1.48	2.78
1971	1.51	1.35	1.45	1.77
1972	1.19	.83	1.11	1.30
1973	1.00	1.11	1.22	—

[a] Supplied by New York City Department of Air Resources.

The "M" step of the algorithm is then calculated in the standard manner of maximum likelihood estimation, assuming y_i and y_i^2 were known. Since there are n_i. observations at each level of X, the r, s element of the X^TX matrix is

$$(X^TX)_{r,s} = \sum_{i=1}^{N} n_{i\cdot}x_{ir}x_{is} \ . \tag{3.9}$$

The rth element of the X^TY vector is

$$X^TY_r = \sum_{i=1}^{N}\sum_{j=1}^{J} n_{ij}X_{ir}\hat{E}(y_i | c_j < y_i < c_{j+1}) \ . \tag{3.10}$$

The quantity Y^TY is

$$Y^TY = \sum_{i=1}^{N}\sum_{j=1}^{J} n_{ij}\hat{E}(y_i^2 | c_j < y_i < c_{j+1}) \ . \tag{3.11}$$

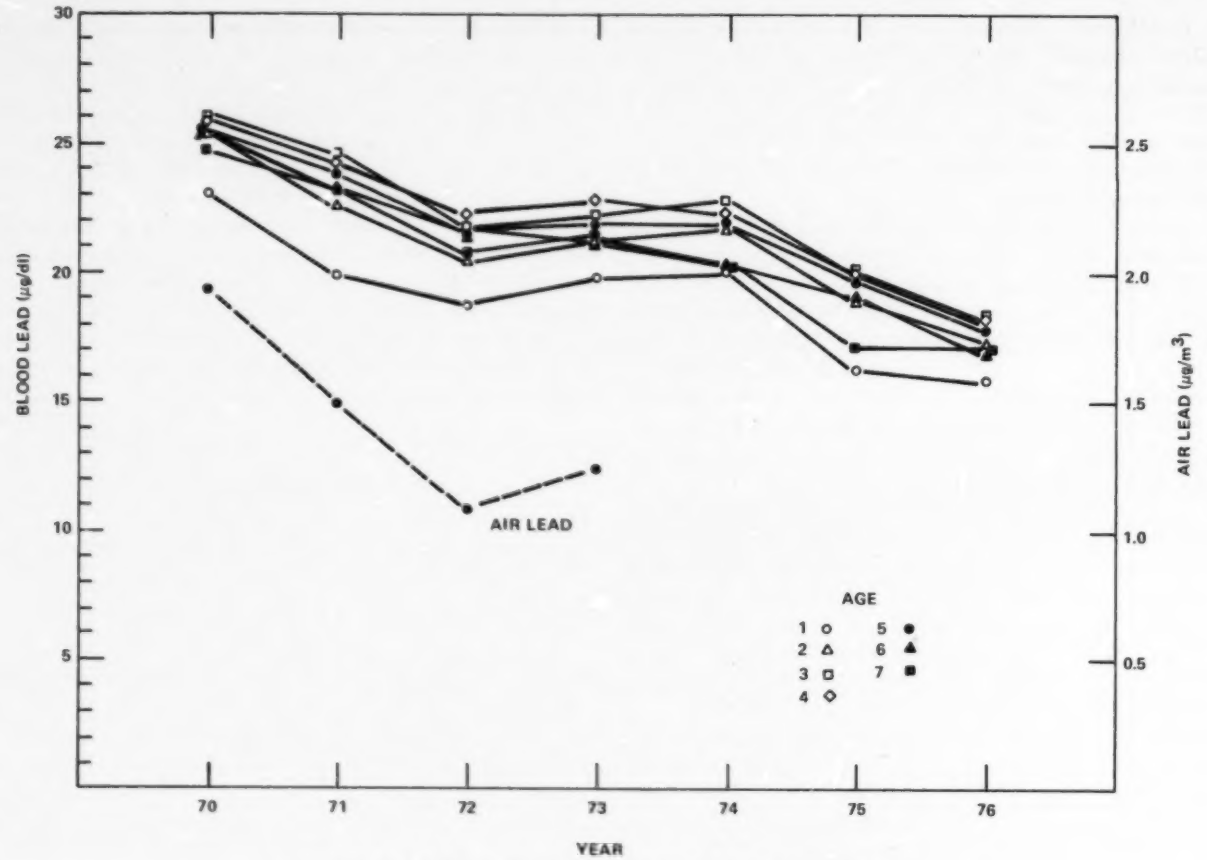
The new estimates of β are

$$\hat{\beta} = (X^TX)^{-1}(X^TY) \ , \tag{3.12}$$

and σ^2 is estimated by

$$\hat{\sigma}^2 = [Y^TY - \hat{\beta}(X^TY)]/N \ . \tag{3.13}$$

Note that the matrix $(X^TX)^{-1}$ need not be recomputed at each iteration.



C. New York City Mean Blood Levels for Puerto Ricans

4. RESULTS OF THE STATISTICAL ANALYSIS

The algorithm defined in Section 3 was applied to our data set. The coded blood lead values were fitted with various models to estimate the effects of the three most representative ethnic groups, age, and the year of the sample's collection. Examining only three ethnic groups resulted in a reduced sample of 130,865 blood lead values (Table 2).

4. Estimates of the Log-likelihood Function for Various Models

Model	Number of Parameters	Log-likelihood
Full model (all-means model)	148	−171268.13
No three-way interaction	76	−171309.17
Main effects plus year × race and year × age	64	−171488.70
Main effects plus year × race	52	−171563.83
Main effects only	16	−171678.49
Age, race, and linear year	11	−172009.51
Age and race only	10	−176231.59
Race, linear age, quadratic age	6	−176402.79
Race and linear age	5	−176996.10
Race only	4	−177012.80
Mean only	2	−179363.57

Estimate of the geometric standard deviation using the full model = 1.42.

A model with each cell fitted by its own mean (all-means model required only 11 iterations to converge and yielded estimates of 24.093 μg/100 ml for the overall geometric mean and 1.421 μg/100 ml for the pooled geometric standard deviation. The estimated individual cell means are shown in Figures A through C. These figures show a fairly consistent trend over years for all age and race groups.

5. One Possible Way of Arranging Differences of Log-likelihood Functions in Table 4 to Form Table of Chi-Squares

Factor	Degrees of Freedom	Asymptotic Chi-square	P value
Three-way interaction	72	82.08	.195
Age × race interaction	12	359.06	<.001
Year × age interaction	12	150.26	<.001
Year × race interaction	12	229.32	<.001
Year	6	9106.20	<.001
nonlinear	5	662.04	<.001
linear	1	8444.16	<.001
Age	6	1562.42	<.001
higher order	4	342.40	<.001
quadratic	1	1186.62	<.001
linear	1	33.40	<.001
Race	2	4701.54	<.001

It is difficult to relate the blood lead means to air lead values because of the lack of air lead readings. Air lead averages for the years 1969 to 1973 are shown in Table 3. These values represent a five-station average for Manhattan. The estimated annual means are plotted on Figures A through C and are consistent with the decrease in blood lead means.

As tests of significance can be obtained by using log-likelihood ratio tests, log-likelihood functions were estimated for several models. The selected models and the corresponding estimates of the log-likelihood appear in Table 4. Multiplying appropriate differences in these functions by -2, the resultant chi-squares can be presented in a fashion similar to the sums-of-squares breakdown in an analysis of variance table. One such table is presented in Table 5. (All of these models required less than 12 iterations to converge.)

Table 5 shows huge main effects, smaller but significant first-order interactions, and no strong evidence of a three-way interaction. The analysis of three ethnic groups showed strong main effects consistent with the results of Billick, Curran, and Shier (1979). In particular, black children had much larger mean blood lead levels than Puerto Rican children, who, in turn, had slightly larger mean levels than white children. The age trend was similar to that found by Yankel, von Lindern, and Walter (1977), with mean blood lead levels increasing up to age three, and then decreasing from age four. This is consistent with the chi-squares for age, which have the largest component in the quadratic term. Of special interest was the sharp decrease in annual mean blood lead levels from 1970 to 1976. This decrease appeared in all age and race subgroups. Much of the decrease is linear, and the graphs indicate that the annual trends show the same trend as the very limited air lead data.

5. DISCUSSION

Some justification for our uneasiness with the simple use of midpoints was borne out by our analysis. For example, the midpoint estimate of the geometric mean for whites age 37 to 48 months in 1974 was 20.1 compared with 13.9 in 1975. Maximum likelihood estimates of the geometric mean for the same periods were 21.1 and 20.1, respectively. Thus the change in interval definition from 1974 to 1975 affected the bias using estimates based on interval midpoints.

A similar effect was observed in the estimate of the geometric standard deviation. The midpoint estimates for the geometric standard deviation for the population cited in the previous paragraph were 1.59 in 1974 and 1.63 in 1975. The corresponding maximum likelihood estimates were 1.40 and 1.44, which are similar to the pooled value of 1.42 for the entire data set.

Although the analysis described here was developed specifically to handle blood lead data, it should also be pointed out that the method need not be restricted to grouped data following a lognormal distribution. In fact, the method could be extended to any set of grouped data, which, on using some suitable transformation, could be assumed normal.

[*Received December 1978. Revised June 1980.*]

REFERENCES

Azar, A., Snee, R.D., and Habibi, K. (1975), "An Epidemiologic Approach to Community Air Lead Exposure Using Personal Air Samplers," *Environmental Quality and Safety, Supplement II: Lead*, 37, 129–145.

Billick, I.H., Curran, A.S., and Shier, D.R. (1979), "Analysis of Pediatric Blood Lead Levels in New York for 1970–1976," *Archives of Environmental Health*, 31, 180–190.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 38, 1–22.

Environmental Protection Agency (1978), "National Ambient Air Quality Standard for Lead," *Federal Register*, 43, October 5, 46246–46277.

Gjeddebaek, N.F. (1949), "Contributions to the Study of Grouped Observations. Applications of the Method of Maximum Likelihood in Case of Normally Distributed Observations," *Skandinavisk Aktuarietidskrift*, 32, 135.

Guinee, V.F. (1971), "Lead Poisoning in New York City," *Transactions of New York Academy of Sciences*, 33, 539.

Hasselblad, V., and Nelson, W. (1975), "Additional Analyses of the Seven Cities Lead Study," *Environmental Quality and Safety, Supplement II: Lead*, 147–551.

Lucas, J.M. (1978), "Effect of Analytical Variability on Measurements of Population Blood Lead Levels," presented at the 1978 Annual Meetings of the American Statistical Association, San Diego.

Orchard, T., and Woodbury, M.A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the 6th Berkeley Symposium*, 1, 697–715.

Tepper, L.B., and Levin, L.S. (1975), "A Survey of Air and Population Lead Levels in Selected American Communities," *Environmental Quality and Safety, Supplement II: Lead*, 152–195.

Woodbury, M.A., and Hasselblad, V. (1970), "Maximum Likelihood Estimates of the Variance Covariance Matrix From the Multivariate Normal," *Proceedings of SHARE XXXIV*, 3, 1550–1561.

Yankel, A.J., von Lindern, I., and Walter, S.D. (1977), "The Silver Valley Lead Study. The Relationship Between Childhood Blood Lead Levels and Exposure," *Journal of the Air Pollution Control Association*, 27, 763–767.